

Zero-Shot Learning with Multi-Battery Factor Analysis

Zhong Ji^a, Yuzhong Xie^a, Yanwei Pang^a, Lei Chen^a, Zhongfei Zhang^b

^a*School of Electronic Information Engineering, Tianjin University, Tianjin, 300072, China*

^b*Department of Computer Science, State University of New York, Binghamton, NY 13902, USA*

Abstract

Zero-shot learning (ZSL) extends the conventional image classification technique to a more challenging situation where the test image categories are not seen in the training samples. Most studies on ZSL utilize side information such as attributes or word vectors to bridge the relations between the seen classes and the unseen classes. However, existing approaches on ZSL typically exploit a shared space for each type of side information independently, which cannot make full use of the complementary knowledge of different types of side information. To this end, this paper presents an MBFA-ZSL approach to embed different types of side information as well as the visual feature into one shared space. Specifically, we first develop an algorithm named Multi-Battery Factor Analysis (MBFA) to build a unified semantic space, and then employ multiple types of side information in it to achieve the ZSL. The close-form solution makes MBFA-ZSL simple to implement and efficient to run on large datasets. Extensive experiments on the popular AwA, CUB, and SUN datasets show its significant superiority over the state-of-the-art approaches.

Keywords: Zero-shot learning, Multi-battery factor analysis, Image classification, Attribute, Word vector.

1. Introduction and Related Work

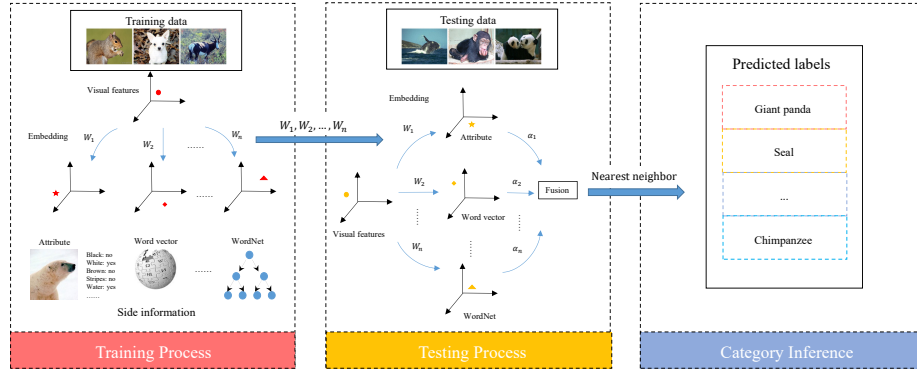
Zero-shot learning (ZSL) aims at solving the problem when the new test image categories are not seen in the training samples [1]. Different from the open set recognition and novelty detection which only distinguish abnormalities in the testing data, ZSL seeks to classify the unseen testing classes [2]. This is a practical problem setting in image classification, as there are thousands of categories of objects we intend to recognize, but only a few of them may have been appropriately annotated. Consequently, it is more challenging than the conventional image classification problems. The key ideas of ZSL are to choose better side information (also known as modalities) and to develop an effective common semantic space. The side information provides a bridge to transfer

knowledge from the seen classes for which we have training data to the unseen classes for which we do not, and the common space offers a fusion feasibility for the visual features and the side information.

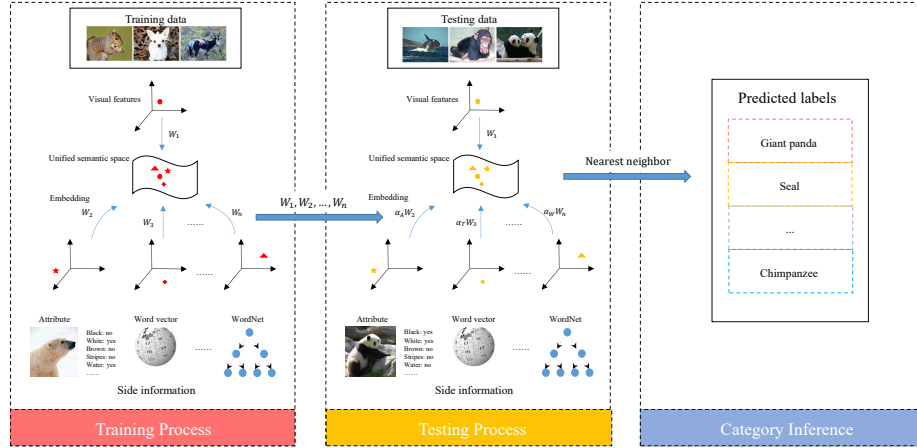
Two types of commonly used side information in ZSL are attributes [3, 4, 5, 6] and word vectors [7], [8]. Particularly, attributes act as intermediate representations shared across multiple classes, indicating the presence or absence of several predefined properties. Direct attribute prediction (DAP) [3] is one of the first efforts to exploit the attributes to ZSL. It learns attribute-specific classifiers with the seen data and infers the unseen class with the learned estimators. However, attribute-based approaches suffer from a poor scalability as the attributes ontology for each class is generally manually defined. Word-vector-based approaches [9, 10, 11, 12] avoid this limitation since word vectors are extracted from a linguistic corpus with neural language models such as GolVe [7] and Word2Vec [8]. Therefore, word vectors have become another popular side information in ZSL. For instance, Socher *et al.* [10] construct a two layer neural network to project images into the word vector space. In [12], Frome *et al.* present a deep visual-semantic embedding model with a hinge loss function, which trains a linear mapping to link the image visual space to the word vector space.

Besides attributes and word vectors, some other side information, such as WordNet [13], visual prototypical concepts [14], class co-occurrence statistics [15], is also applied in ZSL. Further, since different types of side information captures different aspects of the structure of the semantic space, several studies have been made to combine them to achieve higher classification performance [13],[16],[17]. For example, in [13], Akata *et al.* first learn the joint embedding weight matrices corresponding to different types of side information, then perform a grid search over the coefficients on a validation set to get the joint compatibility model. In [16], semantic projections are trained for attributes and word vectors independently, followed by a transductive multi-view semantic embedding space to alleviate the projection domain shift problem. These efforts demonstrate that different types of side information complement each other and construct a better embedding space for knowledge transfer. However, although multiple types of side information are utilized, they still exploit each type of side information in its own semantic space independently, and then just combine the predicted scores together [13], [16]. This cannot make full use of the complementary knowledge of different types of side information. A more efficient and robust solution is to investigate multiple types of side information in a unified space. Unfortunately, to the best of our knowledge, there has been little previous work exploiting this idea. To this end, we present a novel approach called MBFA-ZSL to employ multiple types of side information in a unified space, as shown in Fig. 1.

It is worth highlighting several aspects of the proposed MBFA-ZSL approach. (1) It develops an advanced multi-view embedding algorithm named Multi-Battery Factor Analysis (MBFA), which extends Tucker’s Inter-Battery Factor Analysis (IBFA) [18]. (2) As far as we know, it represents one of the first attempts that embeds both the image visual features and multiple types of side information into one unified semantic space, which fully utilizes the interrela-



(a) Conventional approaches



(b) MBFA-ZSL approach

Figure 1. The comparative illustration of the proposed MBFA-ZSL and the conventional approaches. (a) Conventional approaches embed the visual features to each type of side information in its own semantic space independently, (b) MBFA-ZSL employs multiple types of side information in a unified space.

tions among different types of information. (3) The close-form solution makes it simple to implement and efficient to run on large datasets. (4) Extensive experiments on popular datasets demonstrate its significant superiority over the existing state-of-the-art approaches.

The reminder of this paper is structured as follows. Section 2 introduces the proposed Multi-Battery Factor Analysis (MBFA) algorithm, and Section 3 describes the proposed MBFA-ZSL approach in detail. Experimental results are presented in Section 4, and conclusions are drawn in the final section.

2. Multi-Battery Factor Analysis

Multi-Battery Factor Analysis (MBFA) is developed to provide a unified semantic space for both the visual features and multiple types of side information. It originates from Tucker’s Inter-Battery Factor Analysis (IBFA) [18], which transforms two modalities into a shared space where they are not only well explained but also as much correlated as possible with each other. Thus, we first briefly introduce the IBFA algorithm, and then extend it to a multi-view version, i.e., MBFA.

Given a set of N instances from two modalities, $\mathbf{X}_1 = [\mathbf{x}_{11}, \dots, \mathbf{x}_{1N}] \in \mathbb{R}^{p_1 \times N}$ and $\mathbf{X}_2 = [\mathbf{x}_{21}, \dots, \mathbf{x}_{2N}] \in \mathbb{R}^{p_2 \times N}$, where p_1 and p_2 are their dimensionalities, respectively. With $\mathbf{X}_1, \mathbf{X}_2$ centered, IBFA finds two projection matrices, \mathbf{W}_1 and \mathbf{W}_2 , by the following constrained maximization:

$$\begin{aligned} \max_{\mathbf{W}_1, \mathbf{W}_2} \quad & \text{tr}(\mathbf{W}_1^T \mathbf{X}_1 \mathbf{X}_2^T \mathbf{W}_2), \\ \text{s.t.} \quad & \mathbf{W}_1^T \mathbf{W}_1 = \mathbf{I}, \mathbf{W}_2^T \mathbf{W}_2 = \mathbf{I}. \end{aligned} \quad (1)$$

where \mathbf{I} is an identity matrix. IBFA maximizes the total covariance between the two modalities, which can be seen plainly by rewriting $\text{tr}(\mathbf{W}_1^T \mathbf{X}_1 \mathbf{X}_2^T \mathbf{W}_2)$ as $\sum_{i=1}^N \mathbf{W}_1^T \mathbf{x}_{1i} \mathbf{x}_{2i}^T \mathbf{W}_2$. With the Lagrange multiplier method, (1) can be solved analytically through the eigenvalue decomposition.

Compared with Canonical Correlation Analysis (CCA) [19], (1) can be rewritten as:

$$\begin{aligned} \max_{\mathbf{W}_1, \mathbf{W}_2} \quad & (\text{corr}(\mathbf{W}_1^T \mathbf{X}_1, \mathbf{W}_2^T \mathbf{X}_2) \cdot \sqrt{\text{var}(\mathbf{W}_1^T \mathbf{X}_1)} \cdot \sqrt{\text{var}(\mathbf{W}_2^T \mathbf{X}_2)}), \\ \text{s.t.} \quad & \mathbf{W}_1^T \mathbf{W}_1 = \mathbf{I}, \mathbf{W}_2^T \mathbf{W}_2 = \mathbf{I}. \end{aligned} \quad (2)$$

where $\text{corr}(\mathbf{a}, \mathbf{b})$ denotes the Pearson correlation, and $\text{var}(\mathbf{a}) = \mathbf{a}^T \mathbf{a}$ is the variance. It can be seen from (2) that IBFA attempts to capture both the correlation and variation of \mathbf{X}_1 and \mathbf{X}_2 . This is different from CCA that only aims at maximizing their correlation. In particular, the maximized correlation and variance in (2) depict the relationship between \mathbf{X}_1 and \mathbf{X}_2 and strengthen their own discriminant capabilities, respectively.

To broaden IBFA to a multi-view scenario, we develop the MBFA algorithm on the basis of IBFA. Given a set of N instances from c modalities, $\mathbf{X}_i =$

$[\mathbf{x}_{i1}, \dots, \mathbf{x}_{iN}] \in \mathbb{R}^{p_i \times N}$, $i = 1, \dots, c$, where p_i denotes the dimensionality, with \mathbf{X}_i centered, the objective function of MBFA is expressed as:

$$\begin{aligned} & \max_{\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_c} \sum_{\substack{i,j=1 \\ i \neq j}}^c \text{tr}(\mathbf{W}_i^T \mathbf{X}_i \mathbf{X}_j^T \mathbf{W}_j). \\ \text{s.t.} \quad & \mathbf{W}_i^T \mathbf{W}_i = \mathbf{I}, \quad i = 1, \dots, c, c \geq 2. \end{aligned} \quad (3)$$

Similar to IBFA, MBFA tries to find a set of projection matrices that maximize the total covariance in the common space. Equation (3) can be rewritten as:

$$\begin{aligned} & \max_{\mathbf{W}} \quad \text{tr}(\mathbf{W}^T \mathbf{M} \mathbf{W}), \\ \text{s.t.} \quad & \mathbf{W}^T \mathbf{W} = \mathbf{I}. \end{aligned} \quad (4)$$

where \mathbf{W} and \mathbf{M} are as follows:

$$\mathbf{W} = [\mathbf{W}_1^T, \mathbf{W}_2^T, \dots, \mathbf{W}_c^T]^T. \quad (5)$$

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} & \dots & \mathbf{M}_{1c} \\ \mathbf{M}_{21} & \mathbf{M}_{22} & \dots & \mathbf{M}_{2c} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{M}_{c1} & \mathbf{M}_{c2} & \dots & \mathbf{M}_{cc} \end{bmatrix}, \mathbf{M}_{ij} = \begin{cases} 0, & i = j \\ \mathbf{X}_i \mathbf{X}_j^T, & i \neq j. \end{cases} \quad (6)$$

Equation (4) can be solved via the eigenvalue decomposition; thus each projection matrix \mathbf{W}_i can be obtained. It is obvious that IBFA can be considered as a special case of MBFA when c is 2.

In addition, the main difference between MBFA and Multi-view Canonical Correlation Analysis (MCCA) [20], [21] is worth highlighting. Both of them find a set of linear transformations to project multiple modalities into one common space. However, MCCA seeks to maximize the total correlation in the common space, whereas MBFA maximizes the total covariance, which is equivalent to maximize the total correlation and variance simultaneously. To the best of our knowledge, there is no previous work using MCCA on ZSL. In this paper, we also implement MCCA on ZSL as a comparative approach (we call this approach as MCCA-ZSL).

3. Zero-Shot Learning with MBFA

In a ZSL setting, we are given N_s labeled training instances $\mathcal{S} = \{\mathbf{X}, \mathbf{Y}^k, \mathbf{z}\}$ and N_u unlabeled testing instances $\mathcal{U} = \{\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}^k, \tilde{\mathbf{z}}\}$. $\mathbf{X} \in \mathbb{R}^{p \times N_s}$ and $\tilde{\mathbf{X}} \in \mathbb{R}^{p \times N_u}$ are the p -dimensional visual feature vectors of training and testing instances respectively. \mathbf{z} and $\tilde{\mathbf{z}}$ are the seen and unseen class label vectors, and $\mathbf{z} \cap \tilde{\mathbf{z}} = \emptyset$. We have K different types of side information, $\mathbf{Y}^k \in \mathbb{R}^{q_k \times N_s}$ and $\tilde{\mathbf{Y}}^k \in \mathbb{R}^{q_k \times N_u}$ denote the k -th type of q_k -dimensional side information for training and testing datasets respectively. Note that for the testing dataset, $\tilde{\mathbf{Y}}^k$ is

missing as testing instances are unlabeled. The task of ZSL is to predict the class labels $\tilde{\mathbf{z}}$.

The proposed MBFA-ZSL algorithm mainly contains the following two steps:

Step 1: Build a MBFA space with the seen data. The MBFA algorithm provides an unified semantic space \mathbb{Z} for different types of side information as well as the visual features. With the seen images together with the side information, we can train the MBFA model to obtain a set of projection matrices \mathbf{W}_i ($i = 1, \dots, c$), where c is the sum of all the types of side information and the visual features, such that $c = K + 1$. For example, if we use both attributes and word vectors as the side information, then c is 3.

Step 2: Unseen category Inference. With the projection matrix \mathbf{W}_1 learned from the seen data, the unseen image features $\tilde{\mathbf{x}}_j \in \tilde{\mathbf{X}}$ can be embedded into the common space \mathbb{Z} by $\theta(\tilde{\mathbf{x}}_j) = \mathbf{W}_1^T \tilde{\mathbf{x}}_j$. Typically, the unseen category of $\tilde{\mathbf{x}}_j$ can be inferred by searching for the nearest output embedding vector that corresponds to one of the unseen classes, if there is only single side information available in \mathbb{Z} . Since there are multiple types of side information used in the MBFA-ZSL approach, we introduce a multi-modality fusion method to predict the unseen category of the $\tilde{\mathbf{x}}_j$ with:

$$l^* = \underset{l}{\operatorname{argmax}} \left[\sum_{k=1}^K \alpha_k \operatorname{sim} \left(\theta(\tilde{\mathbf{x}}_j), \varphi_k(\tilde{\mathbf{y}}_l^k) \right) \right], l = 1, 2, \dots, n, \quad (7)$$

where α_k is a weight associated with each type of side information, which can be determined by a grid search on the validation set. Each type of side information that corresponds to the l -th unseen class is denoted as $\tilde{\mathbf{y}}_l^k$, and it can be embedded into the common space \mathbb{Z} by $\varphi_k(\tilde{\mathbf{y}}_l^k) = \mathbf{W}_{k+1}^T \tilde{\mathbf{y}}_l^k$. The similarity between two vectors can be represented as the common distance measurements, such as dot product similarity and Euclidean distance. In this paper, the cosine distance is utilized, i.e., $\operatorname{sim}(\mathbf{a}, \mathbf{b}) = \mathbf{a}^T \mathbf{b} / (\|\mathbf{a}\| \cdot \|\mathbf{b}\|)$

Moreover, MBFA-ZSL has an explicit, close-form solution, which makes it simple to implement and efficient to run on large datasets. **Algorithm 1** outlines the procedures of the proposed MBFA-ZSL approach.

4. Experimental Results and Discussion

4.1. Datasets and Settings

We evaluate the proposed MBFA-ZSL approach on three publicly popular datasets: Animals with Attributes (AwA) [3], Caltech-UCSD-Birds-200-2011 (CUB) [22], and SUN Attribute [23]. Specifically, AwA is a collection of 30,475 images on 50 classes of animals, with 85 associated class-level attributes. We use the standard training/test (seen/unseen) split as that in [3], which chooses 40 classes for training and validation and 10 classes for testing. CUB provides 200 classes of birds (11,788 images), and each class is annotated with 312 attributes. Particularly, CUB is a much more challenging dataset in that it is designed for fine-grained image classification and contains more classes but fewer images.

Algorithm 1 MBFA-ZSL approach

Input: A labeled seen data set $\mathcal{S} = \{\mathbf{X}, \mathbf{Y}^k, \mathbf{z}\}$, an unlabeled unseen data set $\mathcal{U} = \{\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}^k, \tilde{\mathbf{z}}\}$, and the dimensionality d of the unified embedding space.

Output: Labels of the unseen data \mathcal{U} .

- 1: Construct the covariance matrix \mathbf{M} with the labeled visual features \mathbf{X} and the corresponding side information \mathbf{Y}^k .
 - 2: Solve the eigenvalue decomposition problem in (4), and the eigenvectors corresponding to the largest d eigenvalues form the projection matrices \mathbf{W} .
 - 3: Learn the weight parameters of the category inference function (7) in the validation set.
 - 4: Project the unseen visual feature $\tilde{\mathbf{X}}$ and the side information of the unseen classes into the unified space with projection matrices \mathbf{W} .
 - 5: Predict the labels of \mathcal{U} with (7).
-

Similar to [13], we use 150 classes as training and validation set, leaving 50 disjoint classes as test set. SUN Attribute dataset consists of 14,340 images from 717 scene categories, and each category is annotated with a taxonomy of 102 discriminate attributes. We adopt the popular training/test (seen/unseen) split as that in [24], which selects 707 classes for training and validation, and takes the remaining 10 classes as testing set. We cross-validate the parameters α_k in (7). The example images in these datasets are shown in Fig. 2.

On the AwA dataset, we use the VGG (very deep 19-layer CNN) features provided in [25] as visual features. On the CUB and SUN dataset, we use a pre-trained VGG model to extract visual features [26]. For each image, the 4,096 dimensional top-layer hidden unit activations (fc7) of VGG are taken as visual features.

We use both the word vectors (T) and attributes (A) as the side information in MBFA-ZSL. Specifically, we train the Word2Vec model [8] on a corpus of Wikipedia documents to form 1000-D word vectors for the three datasets. Meanwhile, we use the attribute information provided by the datasets. The average per-class top-1 accuracy on the test sets is reported.

4.2. Results on the AwA, CUB, and SUN datasets

We compare the proposed MBFA-ZSL with 7 state-of-the-art approaches as well as MCCA-ZSL, which utilize a range of side information. Among them, DAP [3], [4], ESZSL [6], and SSE-ReLU [27] only use attributes; LatEm [28] can make use of either word vectors or attributes; SJE [13], AMP [29], TMV-HLP [16] and MCCA-ZSL employ more than one type of side information. Different CNN visual features are applied in these approaches, such as GoogLeNet [30], Overfeat [31], and VGGNet-19 [26]. Additionally, we also implement MBFA-ZSL and MCCA-ZSL in the situation where only attributes (A) or word vectors (T) are available.

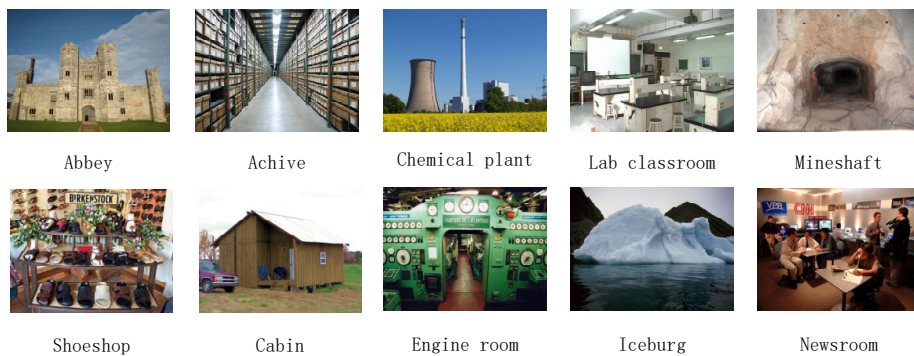
The performance of MBFA-ZSL are taken via ten times of cross validation. It should be noticed that when only T or A is available, the single param-



(a) Example images of the AWA dataset



(b) Example images of the CUB dataset



(c) Example images of the SUN dataset

Figure 2. Example images of the AWA, CUB, and SUN datasets.

ter α_k in (7) does not need to be tuned, thus there is no standard deviation for the corresponding results. Furthermore, the standard deviations of some comparative results are absent as they are not available in the original papers. The comparative results are summarized in Table 1, from which we can observe that MBFA-ZSL achieves the amazingly best performance in all cases for all the three datasets. Besides, we also have the following observations:

Table 1. Performance Comparison (% , mean \pm standard deviation) on the AwA, CUB, and SUN Datasets

| Image feature | Approach | AwA | | | CUB | | | SUN | | |
|---------------|---------------|----------------|----------------|--------------------------------|----------------|----------------|--------------------------------|----------------|----------------|--------------------------------|
| | | T | A | T+A | T | A | T+A | T | A | T+A |
| GoogLeNet-22 | SJE [13] | 51.2 | 66.7 | 73.5* | 28.4 | 50.1 | 51.0* | - | - | - |
| | LatEm [28] | 61.1 | 71.9 | - | 31.8 | 45.5 | - | - | - | - |
| Overfeat-8 | AMP [29] | - | - | 66.0 | - | - | - | - | - | - |
| | TMV-HLP [16] | - | - | 73.5 | - | - | 47.9 | - | - | - |
| VGGNet-19 | DAP [4] | - | 60.8 | - | - | - | - | - | 72.0 | - |
| | SSE-ReLU [27] | - | 76.3 \pm 0.8 | - | - | 30.4 \pm 0.2 | - | - | 82.5 \pm 1.3 | - |
| | ESZSL** [6] | - | 74.6 \pm 3.7 | - | - | 50.8 \pm 0.4 | - | - | 84.5 \pm 1.4 | - |
| | MCCA-ZSL** | 65.8 \pm 1.7 | 74.9 \pm 0.3 | 75.3 \pm 1.8 | 32.1 \pm 0.3 | 45.8 \pm 0.2 | 46.4 \pm 0.7 | 59.5 \pm 1.7 | 82.8 \pm 0.5 | 85.1 \pm 1.5 |
| | MBFA-ZSL** | 72.5 | 77.8 | 79.9\pm0.7 | 32.4 | 51.7 | 52.2\pm0.4 | 61.5 | 85.0 | 87.4\pm0.2 |

T, A represent attributes and word vectors, respectively.

*: additional WordNet hierarchies are used; **: our implementation.

(1) For AwA dataset, the second-best approaches are MCCA-ZSL, SSE-ReLU, and MCCA-ZSL in the cases of T, A, and T+A, respectively. MBFA-ZSL outperforms them in 6.7%, 1.5%, and 4.6%, respectively. For CUB dataset, MBFA-ZSL outperforms the second-best approaches, MCCA-ZSL in 0.3%, ESZSL in 0.9%, and SJE in 1.2% in the three cases, respectively. For SUN dataset, in the three cases, MBFA-ZSL outperforms the second-best approaches, MCCA-ZSL in 2.0%, ESZSL in 0.5%, and MCCA-ZSL in 2.3%, respectively. These are very promising results.

(2) For MBFA-ZSL, the performance on AwA in T+A is better than those in T and A in 7.4% and 2.1%, respectively. On CUB, the performance in T+A is better than those in T and A in 19.8% and 0.5%, respectively. On SUN, the promotions are 25.9% and 2.4%, respectively. Similar observation can also be found in MCCA-ZSL. The excellent performance in T+A of MBFA-ZSL and MCCA-ZSL demonstrates that it is effective to embed multiple types of side information into a unified space. It also confirms that different types of side information complement each other in transferring knowledge.

(3) In the situation of “T+A”, it can be found that MBFA-ZSL outperforms the others significantly. Take the AwA for example, the performance improvements of MBFA-ZSL over SJE, AMP and TMV-HLP are 6.4%, 13.9%, and 6.4%, respectively. This demonstrates that embedding the visual features and multiple types of side information in a unified space is more promising than the conventional methods that projecting the visual features to each type of side information space independently at first, and then combining them together.

(4) When only single type of side information is available, attributes often help achieve a higher accuracy than word vectors. This is due to that attributes are manually defined for a specific dataset, so they are able to describe category relationship of the dataset more effectively; nevertheless, word vectors are extracted from the corpus in an unsupervised manner, whose capacity is con-

strained by the size or specific domain of the corpus, as well as the polysemy issue.

(5) Interestingly, the performance on CUB is inferior to that on AwA and SUN. The reason may lie in the fine-grained characteristic of CUB. Both the visual appearance and the class names in it are similar to each other, which make it hard to recognize.

To clearly evaluate the performance of MBFA-ZSL on each class, we present the confusion matrix of AwA with T+A, as illustrated in Fig. 3. The diagonal elements denotes the correct prediction accuracy of each class, from which we can see that the proposed MBFA-ZSL can achieve relatively high performance on every class.

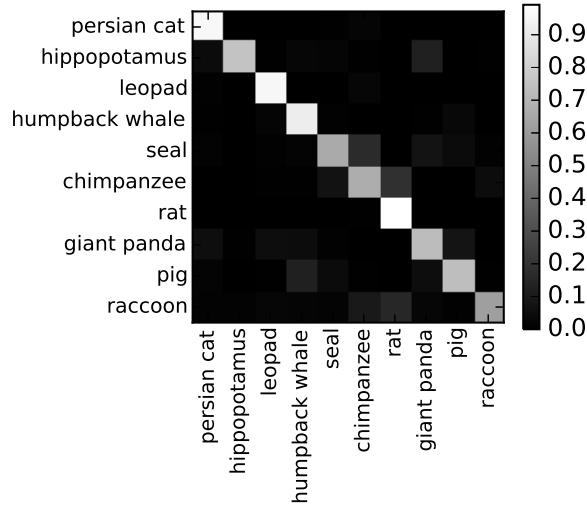


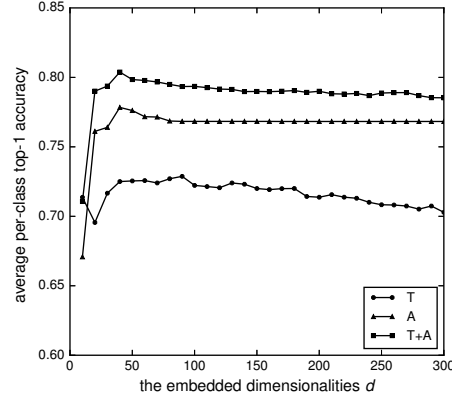
Figure 3. The confusion matrix between test classes of the AwA dataset.

4.3. Parameter Sensitivity

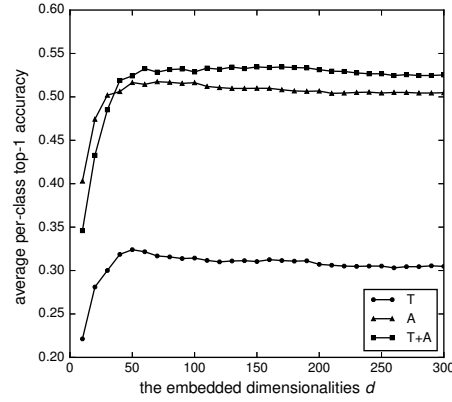
There are two types of parameters in MBFA-ZSL: the weights α_k in (7) and the dimensionality d of the unified space that multiple modalities are projected into. The weights are decided by the cross validation. The impact of dimensionalities d is shown in Fig. 4. The optimal dimensionalities for AwA, CUB, and SUN are 40, 50, and 120, respectively. It can be observed that a higher dimensionality has no performance improvement.

4.4. Speed Evaluation

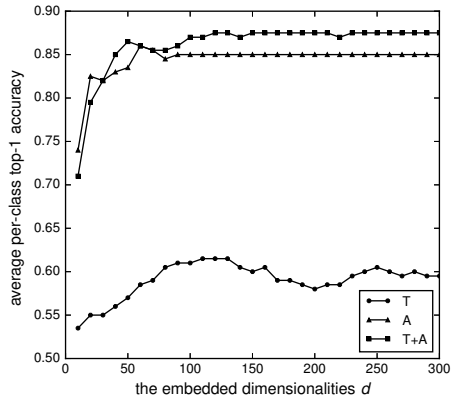
Finally, as shown in Table 2, we report the running times of the training and testing stages for AwA, CUB, and SUN, respectively. Our implementation is based on an unoptimized Matlab code. On our computer with i5 4590 CPU and 12G memory, the training times for the three datasets are 18.8s, 22.2s, and 17.4s, respectively. The test times on each image for the three datasets



(a) The AwA dataset



(b) The CUB dataset



(c) The SUN dataset

Figure 4. The average per-class top-1 accuracy of MBFA-ZSL on AwA, CUB, and SUN datasets with respect to different settings and the embedded dimensionalities d . The types of side information are word vectors (T), attributes (A), and both of them (T+A).

are 0.006ms, 0.010ms, and 0.019ms, respectively. Therefore, MBFA-ZSL is extremely efficient.

Table 2. Training and Testing Times on the AwA, CUB, and SUN Datasets

| | AwA | CUB | SUN |
|---|-------|-------|-------|
| Average training time for all the training data (s) | 18.8 | 22.2 | 17.4 |
| Average training time on each image (ms) | 0.8 | 2.5 | 1.2 |
| Average testing time for all the testing data (ms) | 39.0 | 29.5 | 3.8 |
| Average testing time on each image (ms) | 0.006 | 0.010 | 0.019 |

5. Conclusions

In this paper, we have proposed the MBFA-ZSL approach to projecting both the visual features and multiple types of side information into one unified semantic space to perform ZSL. It can also be applied to the situation where only a single type of side information is available. The results on the three popular datasets show its superior performance over the state-of-the-art approaches on the cases of utilizing A (attributes), T (word vectors), and A+T (attributes + word vectors) as an effective and efficient method. Moreover, it has a close-form solution.

Acknowledgements

This work was supported by the National Basic Research Program of China (973 Program) under Grant 2014CB340400, the National Natural Science Foundation of China under Grant 61271325, Grant 61472273, the Elite scholar Program of Tianjin University under Grant 2015XRG-0014, and the Research Program of Application Foundation and Advanced Technology of Tianjin under Grant 15JCYBJC17100.

References

- [1] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell, “Zero-shot learning with semantic output codes,” in *Advances in Neural Information Processing Systems*, 2009, pp. 1410-1418.
- [2] M. A. F. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, “A review of novelty detection,” *Signal Processing*, vol. 99, pp. 215-249, 2014.
- [3] C. H. Lampert, H. Nickisch, and S. Harmeling, “Learning to detect unseen object classes by between-class attribute transfer,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 951-958.
- [4] C. H. Lampert, H. Nickisch, and S. Harmeling, “Attribute-based classification for zero-shot visual object categorization,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 453-465, 2014.
- [5] J. Deng, N. Ding, Y. Jia, A. Frome, K. Murphy, S. Bengio, Y. Li, H. Neven, and H. Adam, “Large-scale object classification using label relation graphs,” in *European Conference on Computer Vision*, 2014, pp. 48-64.
- [6] B. Paredes and P. Torr, “An embarrassingly simple approach to zero-shot learning,” in *Proceedings of the 32nd International Conference on Machine Learning*, 2015, pp. 2152-2161.
- [7] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the Empirical Methods in Natural Language Processing*, 2014, pp. 1532-1543.
- [8] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Advances in Neural Information Processing Systems*, 2013, pp. 3111-3119.
- [9] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. Corrado, and J. Dean, “Zero-shot learning by convex combination of semantic embeddings,” in *International Conference on Learning Representations*, 2014.
- [10] R. Socher, M. Ganjoo, C. Manning, and A. Ng, “Zero-shot learning through cross-modal transfer,” in *Advances in Neural Information Processing Systems*, 2013, pp. 935-943.
- [11] M. Elhoseiny, B. Saleh, and A. Elgammal. “Write a Classifier: Zero-Shot Learning Using Purely Textual Descriptions,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2584-2591.
- [12] A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, “Devise: A deep visual-semantic embedding model,” in *Advances in Neural Information Processing Systems*, 2013, pp. 2121-2129.

- [13] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele, "Evaluation of Output Embeddings for Fine-Grained Image Classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2927-2936.
- [14] S. Jetley, B. Romera-Paredes, S. Jayasumana, and P. Torr, "Prototypical Priors: From Improving Classification to Zero-Shot Learning," in *British Machine Vision Conference*, 2015, pp. 1-12.
- [15] T. Mensink, E. Gavves, and C. Snoek, "COSTA: Co-occurrence statistics for zero-shot classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2441-2448.
- [16] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong, "Transductive Multi-View Zero-Shot Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 11, pp. 2332-2345, 2015.
- [17] Y. Fu, T. M. Hospedales, T. Xiang, Z. Fu, and S. Gong, "Transductive multi-view embedding for zero-shot recognition and annotation," in *European Conference on Computer Vision*, 2014, pp. 584-599.
- [18] L. R. Tucker, "An inter-battery method of factor analysis," *Psychometrika*, vol. 23, no. 2, pp. 111-136, 1958.
- [19] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, pp. 321-377, 1936.
- [20] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling internet images, tags, and their semantics," *International journal of computer vision*, vol. 106, no. 2, pp. 210-233, 2014.
- [21] J. Rupnik, and J. Shawe-Taylor, "Multi-view canonical correlation analysis," in *Conference on Data Mining and Data Warehouses*, 2010, pp. 1-4.
- [22] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," *Computation & Neural Systems Technical Report*, CNS-TR-2011-001, Caltech, 2011.
- [23] G. Patterson, C. Xu, H. Su, *et al*, "The sun attribute database: Beyond categories for deeper scene understanding," *International Journal of Computer Vision*, vol. 108, no. 1, pp. 59-81, 2014.
- [24] D. Jayaraman, K. Grauman, "Zero-shot recognition with unreliable attributes," *Advances in Neural Information Processing Systems*, 2014, pp. 3464-3472.
- [25] [Online]. Available: <http://attributes.kyb.tuebingen.mpg.de/>
- [26] K. Simonyan and Z. Andrew, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

- [27] Z. Zhang, and V. Saligrama, “Zero-shot learning via semantic similarity embedding,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4166-4174.
- [28] Y. Xian, Z. Akata, G. Sharma, *et al.* “Latent Embeddings for Zero-shot Classification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [29] Z. Fu, T. Xiang, E. Kodirov, and S. Gong, “Zero-shot object recognition by semantic manifold distance,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2635-2644.
- [30] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going Deeper with Convolutions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1-9.
- [31] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, “Overfeat: Integrated recognition, localization and detection using convolutional networks,” in *International Conference on Learning Representations*, 2014.